Introduction:

The competitive space for Deli's (aka Sandwich shops) can be quite dense and saturated with well known, established brands. But let's pretend we have a client that is well versed in everything that has to do with running a sandwich shop. He has everything together, from strategic and operational planning, finances, space renovation skills, equipment and ingredient needs, to having educated guesses as to what menu items will taste the best. Let's say that this client has the hypothesis that if he can penetrate the market in the right location, his brand can capture a lot of positive sentiment in a short amount of time. He is confident that he can operate more efficiently and effectively than many competitors, but wants to maximize the positive shock factor of his grand opening. Therefore, he hires us to find him a new location to open in the city of Toronto. The client also wants to be able to make deliveries within roughly 5 minutes.

Data:

The data used in this project includes Neighborhood postal code data scraped off of Wikipedia, and the GPS coordinates of those neighborhoods as provided by this capstone course. Additionally, the Foursquare API was used to get data about existing venues near certain locations and their ratings. Data was further aggregated or passed into algorithms to create new fields.

Methodology:

The first thing that was done was the scraping of the postal code data of neighborhoods in Toronto. Then using the GPS coordinates of neighborhoods provided to us by the course materials, a combined dataframe with neighborhoods and their coordinates was formed. Neighborhoods were then clustered based on latitude and longitude based on k-means clustering. This means that clusters would be geographically close. The number of clusters was set to 20 because from my judgment, this grouped the neighborhoods into clusters that were, roughly, at most 4 miles apart, give or take a few miles. The closeness of these neighborhoods would allow the client to serve multiple neighborhoods and also make deliveries in relatively short amounts of time. Also k=20 was chosen for the number of clusters because larger values of k would start to result in clusters of single neighborhoods.

After clusters were formed, the GPS coordinates of their centroids were saved and used to create a new dataframe. This dataframe would then hold aggregate data from metrics that were computed off of the Foursquare data. The Foursquare API was queried and the number of venues within 3km of each centroid was computed and stored. The information from the JSONs of venues was kept so that I could then query the API again for ratings of each venue. It should be mentioned that the venues that were searched for were specifically delis, aka sandwich shops. Not just any venue was searched for.

If a rating for a venue was not available at the time, then the rating of such a venue was given a default value of 0 as the assumption that was made was that the venues that had no ratings had inspired no one to rate them with excellent service; and therefore should have a low rating to represent this. Ratings of venues within 3km of each centroid were then averaged and stored with the coordinates of the centroid and the number of venues near the centroid.

Results:

The most opportune centroid was chosen on the basis of which ones had the fewest prospective competitors and which centroid had the lowest average rating. Centroid 19, for cluster 19, was chosen because it had 4 potential competitors, which was relatively low, and none of them had ratings, as was evidenced by an average rating of 0.

Discussion:

As stated before, centroid/cluster 19 was chosen. Upon zooming into the area on the map, it was seen that there were many educational facilities, residential areas, and shopping centers. This would provide ample clientele and also many prospective venue locations to operate from. Given how few delis there were in the area for how densely populated it was, I felt as though it was a great opportunity for the client to open up shop and take the market by storm.

Conclusion:

A little bit a machine learning was used along with some exploratory data analysis. No rigorous use of statistics was done however. But for the scope of the project, and the resources available to the researcher at this moment, the best possible outcome has been chosen in my opinion.